DIGITALCOMMONS
@WAYNESTATE

Journal of Modern Applied Statistical Methods

Volume 3 | Issue 1 Article 24

5-1-2004

A Comparison Of Bayesian And Frequentist Statistics As Applied In A Simple Repeated Measures Example

Jan Perkins
Central Michigan University, jan.perkins@cmich.edu

Daniel Wang
Central Michigan University

Journal of Modern Applied Statistical Methods
May, 2004, Vol. 3, No. 1, 227-233

1538 – 9472/04/$95.00

Early Scholars

A Comparison Of Bayesian And Frequentist Statistics As Applied In A Simple Repeated Measures Example

Jan Perkins
Program in Applied Experimental Psychology
Central Michigan University

Daniel Wang
Department of Mathematics
Central Michigan University

Clinicians see Bayesian and frequentist analysis in published research papers, and need a basic understanding of both. A repeated measures data set was analyzed using both approaches. Assumptions underlying each method and conclusions reached were contrasted. The Bayesian approach is a viable alternative to frequentist statistical analysis for many clinical projects.

Key words: Bayesian statistics; frequentist statistics, clinical research

Introduction

Classical or frequentist statistics is the standard method of analysis in clinical research. There is another statistical option, Bayesian analysis, with advocates arguing that it can be equally or more suited to the analysis of clinical research problems. In recent years increasing numbers of studies have appeared using Bayesian analysis or a combination of Bayesian and frequentist analyses, making it likely that health care clinicians will encounter papers written using a Bayesian approach, and that students will need some exposure to both methods. The purpose of this article is to compare the analysis and interpretation of a simple clinical data set using Bayesian and frequentist approaches as a simplified introduction to the Bayesian approach for clinicians without a background in statistics.

Jan Perkins is a doctoral student in Applied Experimental Psychology, and a Physical Therapy faculty member at Central Michigan University. E-mail at: jan.perkins@cmich.edu. Daniel Wang is an Assistant Professor of Statistics at Central Michigan University. He graduated from the University of Alabama with a Ph.D. in Applied Statistics in 1999.

Methodology

Bayesian analysis has developed from the work of Thomas Bayes, an eighteenth century British Presbyterian minister with an interest in probability theory (Brooks, 2001). His theorem is used in predicting probability. In itself, it is uncontroversial and commonly used in areas such as Mendelian genetics and computerized diagnosis (Lilford & Braunholtz, 1996). For such purposes it is used by statisticians of all backgrounds (Lee, 1989/1992). The application of Bayesian analysis in a broader sense is the source of debate and controversy. An explanation of some of the basic assumptions in these cases may help clarify why there is such heated debate.

Bayesian methods essentially construct probability distributions for unknown quantities of interest given the data, for example the probability that a particular Treatment A is superior to Treatment B given data from a trial. This probability is termed the posterior distribution and then used to reach conclusions about the research question. But in Bayesian analysis researchers are required to estimate a prior distribution for the event of interest in order to run the analysis of a data set. This prior distribution may be based on a variety of external evidence that includes controlled and uncontrolled studies, case reports, and expert opinions. When comparing the two treatments mentioned above, the prior distribution is the probability that Treatment A is superior to Treatment B based on available information before data is collected. The actual data gathered in the study is considered the likelihood.

To state the application of Bayes Theorem in simplistic terms, the posterior probability distribution is proportional to the likelihood of the collected data multiplied by the prior distribution. The likelihood function and the prior function are combined into a distribution summing to 1 to create the posterior probability. All inferences about treatment difference are based on the posterior distribution. With continued data collection, it is possible later to revise the analysis and determine a new (and hopefully more precise) posterior distribution to use in conclusions regarding the superiority of Treatment A. Logically, accumulating evidence would ultimately also change the prior - moving it to a more realistic representation of reality. This updating of the prior distribution occurs as understanding of the phenomenon of interest changes in light of the evidence gathered.

Described in these terms the Bayesian approach has a commonsense appeal: it is possible to give probabilities, integrate information from multiple sources, and revise conclusions in light of new information. The process follows the classical model of scientific thinking and experimentation and is consequently attractive to those trained in the scientific method. Proponents of Bayesian analysis in clinical trials have argued that this makes it flexible and ethical, well suited for subgroup analysis, and offers a good option for ongoing analysis over the course of a trial (Spiegelhalter, Myles, Jones, & Abrams, 1999).

But an acceptable determination of prior distribution is one of the hardest things to do in complex situations, for example when there are conflicting opinions or studies or multiple subgroups to be considered. The incorporation of prior distributions is simultaneously considered the greatest flaw and greatest strength of Bayesian analysis, depending on one's perspective (O'Hagen, Luce & Fryback, 2003; Spiegelhalter et al. 1999). Bayesian calculations have also typically required complex statistical computation power not readily accessible to most clinical researchers.

Bayesian statisticians are working on guidelines for weighting the prior distribution, with skeptical priors being useful if there are important reasons for caution (such as risks or costs of the new treatment), weak priors used when little is known, and optimistic priors being used at selected other times. Guidelines for prior specification are beginning to appear (Kadane, & Wolfson, 1996; Spiegelhalter et al. 1999). It is also possible to use a non-informative or uniform prior which essentially lets the data speak for itself (Box & Taio, 1973; Lee, 1989/1992). The data can of course be analyzed with a variety of priors for subsequent decision making, and indeed data can be collected before knowing the prior, but this demonstrates somewhat sloppy and unscientific thinking. If well done the process should follow the scientific model - the different priors resemble competing hypotheses which are to be tested by examining the data.

The approach in frequentist statistics is philosophically quite different. Probability is viewed as "a limiting ratio in a sequence of repeatable events . . . the ratio becoming ever more exact as the series is extended" (Howie, 2002, p. 1). Data is interpreted using statistical models based on frequencies, with the p-value being a measure of "discrepancy between the data and the null hypothesis" (Goodman, 1999, p. 997). This is very different from the Bayesian view of probability being a degree of belief or knowledge about the unknown. Contrary to common misinterpretations, the p-value does not give the probability of Treatment A being superior to Treatment B, but instead a predetermined level of significance test, set by balancing Type I and Type II errors, that allows acceptance or rejection of the data set based on its compatibility with the null hypothesis. The data are analyzed independently, without the influence of previous knowledge in the analysis, although previous knowledge is of use in planning the data collection. In other words, the classical inference methods treat parameters as constants, while Bayesian methods treat them as random variables.

A frequentist statistician would argue that the introduction of the prior in Bayesian analysis introduces an element of subjectivity that is unacceptable. A Bayesian may counter that the decision to rank Type II errors as less important than Type I and to arbitrarily select a significance level is unscientific. A frequentist statistician may weight multiple tests of variables to reduce the risk of Type I error, selecting a technique for this from a variety of more or less accepted methods. A Bayesian would view an analysis that is more skeptical of Treatment A because you are also looking at other treatments or subgroups as ridiculous. It also could be argued that in frequentist analysis based on sampling, the analysis is only of value if the researcher has chosen the appropriate statistical model and if the data set fits all the assumptions of the chosen model.

If these conditions are not met, classical analysis can act to distort interpretation and the restrictions imposed by the model can exclude relevant information. Goodwin (1999) gave an excellent summary and explanation of issues relating to the use of frequentist and Bayesian analyses in health research.

The result of either type of analysis in uncomplicated situations where model assumptions are similar and where non-informative priors are used often leads to conclusions that are not much different, but at other times this may not be true. It is possible to reach very different conclusions from the same data set. For a general discussion of Bayesian and frequentist statistics with an emphasis on medical research see Matthews (2001a) and related discussion and response (Berger, 2001; Lindley, 2001; Matthews, 2001b; Sasieni, 2001), and an editorial and related articles in the *Annals of Internal Medicine* (Davidoff, 1999; Goodman, 1999). Specific illustrations of how Bayesian analysis can be useful in clinical trials are also readily located (Johns & Anderson, 1999; Lilford, 1999; Simon, 1999).

Problem to be Analyzed

The data set used in this article was generated as part of a student research project. As such it has been analyzed conventionally and prepared for journal submission. This exercise will not give study details but merely use the data set to illustrate Bayesian and frequentist approaches to data analysis and interpretation.

The study examined the short-term effect of a single stretching session on joint range of motion (ROM). Fifteen experimental group subjects were given the treatment (stretch). Measurements were taken at baseline, and at 1, 3,6,15 and 30 minutes post-stretch. Fifteen control group subjects were measured at similar time periods but not subjected to the treatment.

The question of interest was whether the stretching procedure altered the ROM at each of these time points and, if so, whether the stretch altered it more than the process of being measured. It was expected that the six measurements of ROM required in the protocol would affect ROM of the control group, but to a lesser extent. A comparison of the control and experimental groups would therefore be expected to show whether the stretch had any additional effect on ROM. Although Bayesian analysis has analogs to frequentist tests that produce p-values, it was decided to examine 90% confidence intervals and their analogous Bayesian probability intervals.

Results

SPSS for Windows, version 10.1 was used. For each group the baseline was used as an initial reference point with subsequent measures expressed as the difference from this point with the baseline measured being zero. A repeated measures General Linear Model (GLM) analysis was performed with time of measurement as the within subjects factor and group assignment (control vs. experimental) as between subject factor. This analytical model assumes that the measurements are drawn from a normally distributed population and that the different groups have homogeneous variances.

The p-value for testing no difference in the mean change of ROM between the control and experimental groups is 0.000, which leads to the conclusion that there is a difference. Based on the 90% confidence intervals for the estimated mean changes over the time, it is clear that the experimental group performs better than the control group because none of the 90% intervals overlap between groups. These are expressed in Tables 1 and 2, and illustrated graphically in Figure 1.

Table 1: Control group change in measurement from Baseline (Frequentist).

Time of Measurement	Mean	90% Confidence Interval	
		Lower Bound	Upper Bound
Baseline	0	-	-
One minute	1.33	-0.55	3.22
Three minutes	2.27	0.48	4.06
Six minutes	2.67	1.08	4.26
Nine minutes	2.67	0.94	4.39
Fifteen minutes	2.13	0.4	3.86
Thirty minutes	1.87	-0.46	4.2

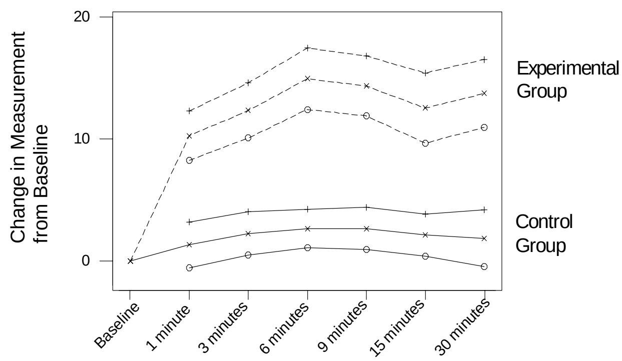
Table 2: Experimental group change in measurement from Baseline (Frequentist).

Time of Measurement	Mean	90% Confidence Interval	
		Lower Bound	Upper Bound
Baseline	0	-	-
One minute	10.27	8.25	12.28
Three minutes	12.33	10.08	14.59
Six minutes	14.93	12.42	17.44
Nine minutes	14.33	11.88	16.79
Fifteen minutes	12.53	9.66	15.4
Thirty minutes	13.73	10.95	16.52

It is possible to state with reasonable confidence that that the data gathered represent the underlying state of affairs. Thus, it could be concluded that the stretch produced an alteration in range of motion that is greater than that caused by the measurement technique.

There is, however, reason to be concerned about the analysis. The statistical model rests on a number of assumptions. If these assumptions are violated there is less faith in the conclusions. It is assumed subjects are a random sample from a pool of suitable subjects and that the raw scores for them (and so the error terms) are normally distributed. It is also assumed that that error terms have a mean of zero and a common variance, and that error terms between and within the groups are not related. These assumptions are based on random assignment of subjects.

Figure 1: 90% Confidence Intervals Using Frequentist Analysis.



In addition there is a complex assumption known as the sphericity assumption related to variances in the fixed factor of the design. The general linear model procedure tests for sphericity using Mauchley's test. In this sample, the test concluded that the assumption was not met. This interpretation was based on using the conservative Greenhouse-Geiser adjustment.

Analysis with Bayesian Statistics

As with most Bayesian analyses, the choice of a prior distribution was critical. In this case there was limited previous evidence to use in creating a prior distribution. Published studies using the particular technique studied did not use the same joints, protocol or exact technique. Clinical experience suggested that there would be a modest increase in range in the experimental group that might or might not decline over the 30 minute period. Experienced clinicians could not offer more specific ideas about the effect of this single stretch treatment.

This limited evidence made it appropriate to use a non-informed prior distribution in the analysis.

The analysis was done using Gibbs sampling, a technique commonly used in Bayesian analysis. Gibbs sampling is a variant of Markov chain Monte Carlo (MCMC) analyses. This computer intensive technique provides researchers with repeated random data points drawn form the statistical distribution of interest. Parameters of interest are estimated by repeated iterations of the process until estimates converge. Gibbs sampling helps compensate for small data sets such as those generated in this experiment. For additional information on the Gibbs sampling technique see Casella and George (1992).

WinBUGS version 1.3 was used in the analysis adapting a dynamic model used in repeated measure research and described in Congdon (2001). The software program is available through the Bayesian inference Using Gibbs Sampling (BUGS) project http://www.mrc-bsu.cam.ac.uk/bugs/). Again, baseline measures were converted to zero and subsequent measures to differences from baseline. Bayesian means and 90% probability intervals were calculated. These are presented in Tables 3 and 4 and Figure 2.

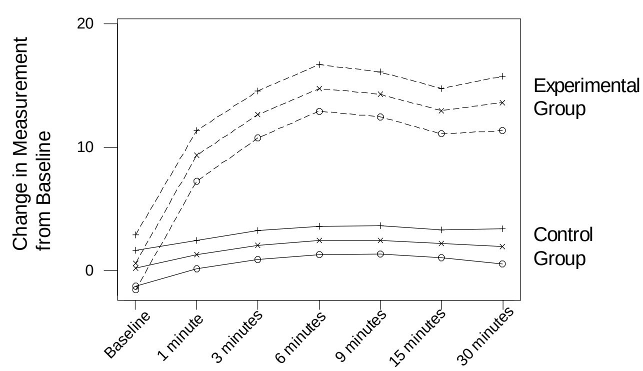
Table 3: Control group change in measurement from Baseline (Bayesian).

Time of Measurement	Mean	90% Probability Interval	
		Lower Bound	Upper Bound
Baseline	0.21	-1.25	1.64
One minute	1.28	0.13	2.47
Three minutes	2.07	0.9	3.26
Six minutes	2.45	1.32	3.62
Nine minutes	2.45	1.37	3.67
Fifteen minutes	2.19	1.04	3.32
Thirty minutes	1.96	0.53	3.38

Table 4: Experimental group change in measurement from Baseline (Bayesian).

Time of Measurement	Mean	90% Probability Interval	
		Lower Bound	Upper Bound
Baseline	0.6	-1.57	2.9
One minute	9.35	7.25	11.36
Three minutes	12.64	10.76	14.56
Six minutes	14.76	12.92	16.7
Nine minutes	14.29	12.47	16.09
Fifteen minutes	12.96	11.08	14.76
Thirty minutes	13.61	11.35	15.73

Figure 2: 90% Probability Intervals Using Bayesian Analysis.



Interpretation was done through examination of the plots and data. Again, there is no overlap in the intervals except at the baseline, where this is expected. The Gibbs sampling technique used produces a baseline estimation, making it possible to give a probability interval for this as well as for the repeated measurement points. The results for estimating the mean change of ROM are very similar to the GLM results but the probability intervals are smaller than the confidence intervals and none of the probability intervals, other than the baseline, contains zero. The Bayesian analysis, like the general linear model, assumes random sampling, normally distributed raw scores for the subjects and a linear relationship between scores and group and time variables. It also makes the important assumption of a non-informative prior.

Comparison of Analyses

In this simple example the conclusions reached with both analytical techniques appear quite similar in terms of clinical interpretation of results and related treatment planning. With this data set and a non-informative prior this is not surprising. Both types of analyses would lead to the practical clinical conclusion that the stretch altered range of motion for at least thirty minutes. In addition there was the expected observation that the repeated measurements did have an effect on ROM, albeit a smaller effect than stretch and measurement combined.

There are, however, some key differences in the interpretation of the results. In the frequentist analysis, the null hypotheses that were no differences in the mean change of ROM is rejected. This conclusion can be reached through the 90% confidence intervals for the mean change without considering any previous information about the mean change. On the other hand, in Bayesian analysis, the distribution of the mean change was estimated and the likelihood of the mean change in terms of the probability intervals calculated. With Bayesian analysis, it is allowed to utilize the previous knowledge about the distribution of parameters.

There are a few differences apparent that may lead to a preference for the Bayesian analysis for this study. The data set is small and does violate some of the assumptions behind the general linear model with repeated measures used in the frequentist analysis. The effect of this is to weaken faith in the conclusions.

The 90% confidence intervals with the general linear model are also wider in all but the one minute measurement in the experimental group analysis than the corresponding Bayesian 90% probability intervals. The width of confidence intervals in conventional analysis gives an estimate of precision with narrower widths desirable (Brooks, 2003). None of the post-baseline Bayesian probability intervals includes zero while two of the frequentist confidence intervals do in the control group, despite an anticipated measurement effect. The smaller intervals in the Bayesian analysis reflect the strength of the sampling procedure and its ability to deal with small data sets. The Bayesian results are more compatible with clinical expectations based on muscle stretching theories. For these reasons the authors conclude that the Bayesian analysis seems to be the better analysis option with this particular data set.

There are additional advantages for a clinician who wants to continue data collection on stretching techniques but lacks facilities for large scale experimentation. The posterior distributions determined from this study could be used as informed priors in subsequent research, refining estimates and improving accuracy with additional data collection. This approach mimics the classic model of scientific reasoning. Assuming the clinician has access to computing resources and programs for Bayesian analysis, a series of small clinical studies could incrementally add to the body of research on the subject. The reasoning process in Bayesian analysis also has its attractions. Ashby and Smith (2000) argue that the Bayesian approach is the natural one for use in evidence-based practice where information must be synthesized and used in individual decision-making.

Conclusion

As computing power increases and statistical packages become more readily available and usable, Bayesian analysis may be seen more often in the medical and health literature used to guide clinical practice. It is now not uncommon to see articles in clinical journals that use Bayesian analysis either alone or in combination with frequentist analysis. This article gives an illustration of Bayesian and classical analysis applied to a simple clinical problem and the interpretation of results. In the example used, the authors concluded that they would prefer the Bayesian approach for analysis. Future studies such as simulating the power of two types of analysis in detecting the mean change of ROM would help clinicians understand the advantage of using classical statistics and Bayesian statistics.

Whatever approach is used in data analysis, it is important to recognize that there is more than one approach. Bayesian analysis is being used in clinical studies to guide practice. In this paper Bayesian and frequentist statistical approaches are used to analyze a sample data set in order to contrast the two approaches and make clinicians aware of different approaches to data analysis.

References

Ashby, D., & Smith, A. F. (2000). Evidence based medicine as Bayesian decision-making. *Statistics in Medicine*, *19,* 3291-3205.

Berger J. (2001). Why should clinicians care about Bayesian methods. *Journal of Statistical Planning and Inference*, *94*, 65-67.

Box, G. E. P., & Taio, G. C. (1973). *Bayesian inference in statistical analysis.* Reading, MA: Addison-Wesley.

Brooks, E. B. (2001). Thomas Bayes. Retrieved November 5, 2003, from http://www.umass.edu/wsp/acquiring/tales/bayes.html

Brooks, G. (2003). Advantages of confidence intervals in clinical research. *Cardiopulmonary Physical Therapy Journal*, *14* (3), 12-14.

Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *American Statistician*, *46,* 167-174.

Congdon, P. (2001). *Bayesian statistical modeling*. New York: John Wiley & Sons.

Davidoff, F. (1999). Standing statistics right side up. *Annals of Internal Medicine, 130,* 1019-1021.

Goodman, S. N. (1999). Toward evidence-based medical statistics. 1: The *P* value fallacy. *Annals of Internal Medicine, 130*, 995-1004.

Goodman S. N. (1999b). Toward evidence-based medical statistics. The Bayes factor. *Annals of Internal Medicine, 130,* 1005-1013.

Howie, D. (2002). *Interpreting probability. Controversies and development in the early twentieth century*. Cambridge, England: Cambridge University.

Johns, D., & Anderson, J. S. (1999). Use of predictive probabilities in Phase II and Phase III clinical trials. *Journal of Biopharmaceutical Statistics*, *9*, 67-79.

Kadane, J. B., & Wolfson, L. J. (1996). Priors for the design and analysis of clinical trials. In D. A. Berry & D. K. Stangl (Eds.), *Bayesian biostatistics* (pp. 157-184). New York: Marcel Dekker.

Lee, P. M. (1989/1992). *Bayesian statistics: An introduction.* New York: Halsted Press.

Lilford, R. J. (1999). Clinical trials and rare diseases: a way out of a conundrum. *BMJ*, *311,* 1621-1625.

Lilford, R. J., & Braunholtz D. (1996). For debate: The statistical basis of public policy: A paradigm shift is overdue. *BMJ*, *313,* 603-607.

Lindley, D. V. (2001). Why *should* clinicians care about Bayesian methods. *Journal of Statistical Planning and Inference, 94*, 59-60.

Matthews, R. A. J. (2001a). Why *should* clinicians care about Bayesian methods. *Journal of Statistical Planning and Inference, 94*, 43-58.

Matthews, R. A. J. (2001b). Author's response *Journal of Statistical Planning and Inference*, *94*, 69-71.

O'Hagen, A., Luce B. R., & Fryback, D. G. (2003). *A primer on Bayesian statistics in health economics and outcomes research.* Bethesda, MD: Medtab International.

Sasieni, P. (2001). Why *should* clinicians care about Bayesian methods. *Journal of Statistical Planning and Inference, 94*, 61-63.

Simon, R. (1999). Bayesian design and analysis of active control clinical trials. *Biometrics*, *55*, 484-487.

Spiegelhalter, D. J., Myles, J. P., Jones, D. R., & Abrams, K. R. (1999). An introduction to Bayesian methods in health technology assessment. *BMJ*, *319,* 508-512.